Census 2000 Testing, Experimentation,
and Evaluation Program September 25, 2003

Topic Report Series, No. 7

Data Processing in Census 2000

FINAL REPORT

This topic report integrates findings and provides context and background for interpretation of results from Census 2000 evaluations, tests, and other research undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation program, designed to assess Census 2000 and to inform 2010 Census planning.

Nicholas Alberti
Decennial Statistical Studies Division

USCENSUSBUREAU
Helping You Make Informed Decisions

The Census 2000 Testing, Experimentation, and Evaluation Program provides measures of effectiveness for the Census 2000 design, operations, systems, and processes and provides information on the value of new or different methodologies. The results and recommendations from these analyses provide valuable information crucial to planning the 2010 Census. By providing measures of how well Census 2000 was conducted, this program fully supports the Census Bureau's strategy to integrate the 2010 early planning process with ongoing Master Address File/TIGER enhancements and the annual American Community Survey. The purpose of the report that follows is to synthesize results from Census 2000 evaluations, experiments, and other assessments and to make recommendations for planning the 2010 Census. Census 2000 Testing, Experimentation, and Evaluation reports are available on the Census Bureau's Internet site at: http://www.census.gov/pred/www/.

TABLE OF CONTENTS

LIST OF TABLES

1. BACKGROUND

1.1 Introduction

The Data Processing Topic Report provides a synthesis of the results, recommendations and lessons learned from the Census 2000 post-data capture data processing. The Census 2000 incorporated varied methods of data collection and data capture. Once census data were collected and captured a collection of interdependent processes were implemented to organize and integrate the data. These processes accomplished the tasks of organizing and integrating individual responses to the census, editing and coding data, integrating geographic data with census response data, identifying and geocoding addresses added through enumeration activities, identifying the best data to represent each household and group quarters, determining the final Census housing universe and ultimately the Census population based on all available data.

1.2 Data Processing Background

1.2.1 Non-ID Processing

Most of the information in this section comes from Medina (2001).

Every address in the census has a unique identifier, the Master Address File (MAF) identification (ID) number. This ID number is used to link each census response to its address. Most census addresses are assigned an ID number prior to the census enumeration operations and most census responses had a MAF ID preprinted on the questionnaire. However, some operations generated responses without a preassigned MAF ID. The Non-ID operation attempted to assign a MAF ID to those responses.

Response records without a MAF ID were divided into three groups. A description of these groups and the Non-ID processes follows:

-- Type A Records - This group included housing unit addresses for responses from the Be Counted program, the Telephone Questionnaire Assistance (TQA) operation and Service-Based Enumeration (SBE) operation, Usual Home Elsewhere (UHE)[1] addresses provided on Group Quarters (GQ) questionnaires (GQ/UHE addresses) and UHE addresses provided on enumerator questionnaire responses in response to the In-Mover and Whole Household UHE coverage improvement probes.

-- Type B Records - This group included responses, from the Be Counted program, that indicated the respondent had no usual home on April 1, 2000. These responses could be included in the GQ universe if the Geography Division (GEO) identified the Local Census Office (LCO) geography that contained the address.

[1] A Usual Home Elsewhere address is a Census Day address reported by a respondent when it is different from the address at which they are interviewed.

-- Type C Records - This group included housing unit addresses that were added to the census through the Update/Leave (U/L), Urban Update/Leave (UU/L), Nonresponse Followup (NRFU), Coverage Improvement Followup (CIFU), Transient-night (T-night), or GQ enumeration programs.

The Decennial Systems and Contracts Management Office (DSCMO) identified the Non-ID records and forwarded them to the GEO for processing. The GEO provided a Census ID number (MAF ID) for each address it could either match or geocode. It updated the MAF with any new housing unit addresses found among the Non-ID responses. The GEO forwarded the results of the Non-ID process to the DSCMO who added the new addresses to the Decennial Master Address File (DMAF).

Type A Record Processing

The GEO conducted an automated match of city style (i.e., house number and street name) and non-city style addresses to the MAF. It also carried out an automated process to geocode city-style addresses that could not be matched to the MAF in the automated process.

The GEO also carried out an interactive telephone and computer assisted operation in National Processing Center (NPC) to match and geocode records that could not be matched or geocoded in the automated processes. If the initial attempt to clerically match or geocode an address failed, the address was compared to a commercial database of addresses in order to obtain a telephone number and/or correct any deficiencies in the address. If appropriate, a second attempt was made to clerically match or geocode the address based on updated address information. If still unsuccessful, the clerical staff used the telephone number to contact the respondent and correct any errors in the address information. If corrections were made, another attempt was then made to match or geocode the address.

New addresses (i.e., those not matched to the MAF) that could be geocoded were added to the Decennial Master Address File (DMAF). Census plans specified that existence of new Type A addresses added to the DMAF through the Non-ID process would be confirmed by the Field Verification (FV) operation. Enumerators visited the location of the new addresses in the FV operation to determine whether or not the address existed as a Census housing unit on April 1, 2000.

Type B Record Processing

The Type B addresses were geocoded only to the geographic area of the LCO since only geographic information collected was the place and county where the person without a usual residence stayed on Census Day. New Type B address locations geocoded to the LCO geography were added to the DMAF.

Type C Record Processing

The GEO assigned a MAF ID to all Type C addresses. The GEO attempted to first match the address to an existing address on the MAF. If no match was found and it could be geocoded the address was added to the DMAF.

1.2.2 Decennial Response File Processing

The Decennial Response File (DRF) processing merged, organized and edited various data response files produced from the paper data capture processes, and Internet and telephone data collection processes. Each response for a Census address was represented on the DRF by a return level record (housing unit level data). The DRF contained one person level record for each person reported on census questionnaires each of which was linked to the appropriate return level record. There could be more than one response for an address and thus more than one return level record and associated set of person level records.

The DRF processing encompassed the following major tasks: merging response data from the data capture output databases with the DMAF to create an initial database of in-scope responses, reformatting and editing the in-scope response data, assigning a housing unit status to each return level record and selecting the response data that would be accepted as the response for each address in subsequent processing.

All of this was carried out in two phases. Each phase is discussed in a separate section below.

1.2.2.1 Decennial Response File - Phase 1 (DRF1)

Much of the information in this section comes from Fowler (2003).

The Decennial Response File - Phase 1 (DRF1) process included the following functions:

-- standardizing data formats
-- merging data capture response data with the DMAF to create an initial file of in-scope responses
-- interfacing with the Non-ID process and DMAF updates.
-- interfacing with the Automated General Coding operations and the Coverage Edit Followup operation.
-- initial editing of the response data to blank illegal characters
-- identifying data defined person records[2] and formatting generational name suffixes

[2] Data defined person records are records with at least minimum data for two or more of the 100 percent census data items - name, gender, relationship, age/date of birth, race and Hispanic origin.

The primary inputs to the DRF1 processing were the Decennial Capture System 2000 (DCS2000) output files of data capture response records transmitted to the DSCMO. The DRF input files also included data from the automated TQA response file and response data received via Internet responses. These files contained records for 82 different types of questionnaires in 15 different formats. The data on these files were reformatted into one standard format for all questionnaire types. This process created the normalized data capture files. The normalized data capture files were then merged with the DMAF. The product of this merge was the initial DRF1, a database consisting of a collection of 559 files containing data response records for census housing units. There was one DRF1 file for each LCO. Only data response records for addresses represented on the DMAF at the time of the merge operation were included on the DRF1.

1.2.2.2 Decennial Response File - Phase 2 (DRF2)

Much of the information in this section comes from Rosenthal (2003).

The second stage of the DRF processing consisted of the following steps:

-- reformatting data
-- linking response records (e.g., linking response records representing enumerator forms with those representing the corresponding continuation forms)
-- determining the housing unit status and household size implied by the data on each housing unit return
-- running the Primary Selection Algorithm process which selected the most appropriate housing unit return(s) for each address

Linking Response Records

The linking of response records refers to the process of combining response records to form housing unit returns. Each housing unit return could be made up of one or more response records (i.e., the data capture records for census forms). When response records were linked, the resulting housing unit return included the demographic data for all persons from the linked response records. One response record was identified as the parent record which contributed the housing unit level data to the housing unit return that was formed.

This linking component of the DRF2 was primarily aimed at linking the response records for the Simplified Enumerator Questionnaire (Forms D-1(E), D-2(E)) with the response records for enumerator continuation forms (Forms D-1(E)Supp and D-2(E)Supp). The Simplified Enumerator Questionnaire provided space to list data for five persons. At a household with six or more persons, an enumerator used a continuation form to complete the enumeration of persons in the household.

The control information on this continuation form identified the Census address for which the questionnaire was associated but did not indicate to which questionnaire it was associated.

When there was only one response for an address the linking of the response record for enumerator questionnaires to the response record of the appropriate continuation form was a simple matter. When there were response records for two or more enumerator questionnaires and one or more continuation forms the linking was not trivial. We used information about the household size reported on the enumerator questionnaires and the number of persons enumerated on the continuation form(s) to identify the most likely linkage between response records.

We also designed the linking process to handle cases where enumerators and mail respondents used questionnaires other than a continuation form to complete the enumeration of a large household and cases where the enumerators used the continuation form to complete the enumeration begun on a mail return. Criteria similar to those used to link response records for enumerator questionnaires to continuation forms were used to link questionnaires in these other types of cases.

When two response records were linked on the DRF2, the questionnaire to which the continuation form was linked was designated as the parent record. The response record for the continuation form was deleted from the file and the control information on the associated person records was updated so that the person records would be associated with the parent record.

Assigning Housing Unit Status and Household Size

Once the DRF2 linking process was completed, the housing unit status and household size of each return was assigned. The information used to assign the housing unit status and household size included the number of persons records associated with a return, the number of names listed in the questionnaire roster, the household size reported by the respondent, and the occupancy status and household size summary information completed by Census enumerators. These data were compared within a return for consistency and examined for sufficiency.

If the data were sufficient and reasonably consistent, the housing unit status and household size were resolved according to a pre-specified set of rules. Small inconsistencies in the data were permitted as long as there was convincing evidence of the status and household size. In each case the status was supported by more than one of the response data items. For example, the number of persons enumerated on a mail return could differ from the household size reported by the respondent. The housing unit is obviously occupied. If in this case, the household size reported by the respondent was five or less, the household size assigned was the maximum of the values indicated by the questionnaire response data.

Each return was given one of the following six occupancy statuses: Occupied, Vacant, Delete, Occupied/Unresolved Household Size, Occupancy Status Unknown and Status Undetermined.

The Occupied and Vacant statuses were assigned when the return clearly showed that the address was an occupied or vacant housing unit. If the data clearly showed that the address was not a Census housing unit or was nonexistent, the Delete status was assigned.

Each housing unit return assigned a status of Occupied was assigned a household size based on all available information including the total number of person records associated with the return, the number of names on the questionnaire roster, the respondent reported household size, and the household size reported by an enumerator in the Interviewer Summary section of the Simplified Enumerator Questionnaire.

The latter three statuses were assigned whenever the housing unit return contained insufficient or conflicting data about housing unit status. They indicate the various levels of an unresolved housing status and/or household size.

-- Occupied with Unresolved Household Size - This status was assigned whenever the response data indicated the address was occupied but the information on household size was insufficient.

-- Occupancy Status Unknown - This status was assigned whenever the information on household status indicated that the address was a Census housing unit but due to conflicts or deficiencies in the response data we could not determine whether or not the housing unit was occupied or vacant.

-- Status Undetermined - This status was assigned whenever we could not determine whether or not the address was a housing unit because the response data provided extremely conflicting information or were extremely deficient.

Primary Selection Algorithm

Most of the information in this section comes from Baumgardner (2002)

The Primary Selection Algorithm (PSA) process was designed to resolve the receipt of multiple responses to the Census 2000 for an individual housing unit address. More than one response could be received for an address because there were several ways to respond to the Census 2000. These included responding via mail, responding via the Internet or telephone, completing a BCF, or being enumerated by a census enumerator as part of the List Enumerate (LE), the NRFU, CIFU or GQ enumeration operations.

It operated on housing unit returns after the housing unit status and household size were assigned. When the DRF2 contained multiple housing unit returns for a census address, the PSA selected the housing unit return(s) along with the associated the set(s) of person level records that best represented the household at that address.

The PSA formed PSA households by combining housing unit returns which represented the same census household. When multiple returns were present for a single address, the PSA matched the names of household members across the returns (within an address) to determine which responses represented the same household. Names and demographic characteristics were used to identify duplicate records for the same person. The presence of duplicate household members

across two returns was evidence that two census responses were completed for the same household. When duplicate household members were identified, the two housing unit returns were combined to form one PSA household.

A PSA household consisted of just one return if there was only one return for the address or if a return had no household members in common with any other returns at the address. It consisted of two or more returns if matching person records were found across the returns. All returns with a status of vacant were combined to form one PSA household.

When a PSA household was formed from two or more returns, one of the returns was designated as the Basic Return. All of the data on the Basic Return were associated with the PSA household while only the demographic data for selected household members from the other return(s) were associated with the PSA household in subsequent data processing.

Once the PSA households were formed, selection criteria were applied to all PSA households at an address to identify the most appropriate PSA household to represent the enumeration at that address. The selected PSA household was designated as the Primary PSA Household. At each address, persons from questionnaires with a Respondent Provided Address (RPA)[3] could be added to the Primary PSA Household if they were not matched in the earlier steps of the PSA. The data that made up the Primary PSA Household were then used as the input to the Hundred Percent Census Unedited File (HCUF) and other subsequent data processing activities.

1.2.3 Hundred Percent Census Unedited File for Housing Units

Much of the information in this section comes from Jonas (2003).

The processing of housing unit data and group quarters data was conducted independently on parallel tracks. The HCUF processing of housing unit data is discussed below. The processing of group quarter data is discussed separately in Section 1.2.4, below. Once both processes were completed, the data were combined to form the final HCUF.

The preliminary Census housing unit universe was determined through the HCUF process applied to potential housing unit addresses on the DMAF. The HCUF housing unit universe was preliminary because it contained records representing duplicate housing units. Potential duplicate HCUF records were identified and retained on the HCUF. The final determine of duplicate records was completed after the HCUF processing. The duplicate records were subsequently removed during of the processing of the Hundred Percent Census Edited File (HCEF).

The HCUF process for housing unit data brought together the data from the DMAF and the

---

[3] BCF's, and GQ questionnaires with a GQ/UHE address.

DRF2. The process determined which addresses on the DMAF represented a census housing unit. The occupancy status of each housing unit was then assigned and the size of the household in each occupied unit was determined.

1.2.3.1 Identifying 'Kills'

The first step in the HCUF processing for housing unit data was to identify addresses on the DMAF that do not represent a housing unit. The addresses eliminated from the housing unit universe at this stage of processing were referred to as 'Kills'. The determination of 'Kills' was based on source of addresses, status of addresses in the U.S. Postal Service Delivery Sequence Files, results of the Local Update of Census Addresses (LUCA), the postal delivery statuses and results of field address listing and housing unit enumeration operations.

1.2.3.2 Integrating the DMAF and DRF2

The next step in the HCUF process was to bring together data for the remaining DMAF addresses and DRF2 data. At this stage of the process one of the following housing unit statuses was assigned to each address on the DMAF: Occupied, Vacant, Delete, Occupied with Unresolved Household Size, Occupancy Status Unknown, and Status Unknown. These statuses correspond to those assigned to individual returns during the DRF2 processing. The latter three categories represent addresses with incomplete information on housing status and/or household size.

The source of the data that determined the housing status and household size of occupied units could be the status and household size assigned to the DRF2 housing unit return selected by the PSA as the Primary PSA Household, or the results of the Nonresponse Followup, the Coverage Improvement Followup or the Field Verification Followup operations as recorded on the DMAF.

1.2.3.3 Count Imputation

The count imputation process imputed the housing status and/or household size to addresses as necessary in the next step. Imputation processes were conducted independently for the three groups of addresses with incomplete information.

-- Occupied with Unresolved Household Size (Household Size Imputation) - A household size of one or greater was imputed to housing units at these addresses.

-- Occupancy Status Unknown (Occupancy Imputation) - A housing unit status of Occupied or Vacant was imputed to these addresses. Addresses in the Occupancy Status Unknown category were determined to exist as a housing unit but it could not be determined if the unit was occupied or vacant. A household size was imputed to households given an imputed status of Occupied.

-- Status Unknown (Status Imputation) - The imputed status of these units could be

Occupied, Vacant or Delete. Status Unknown was assigned whenever we could not determine whether or not the address existed as a Census housing unit because of the lack of sufficient data. Addresses assigned an imputed status of delete were eliminated from the HCUF.

The imputation method used to impute housing status and household size to addresses from each group was a nearest neighbor hot deck imputation method. Donor households were identified for each group from among the addresses with a resolved status and household size. The nearest neighbor housing unit in the donor pool was selected to fill in the housing status and/or household size for the address with incomplete information. The imputation for each group was carried out independently within each LCO.

1.2.3.4 Duplicate Delete Operation

The final step in the HCUF process for housing units was the identification of duplicate addresses. The operations to identify duplicate addresses were designed and conducted in the summer and fall of 2000 to correct a potential overcount of housing units. Two primary methods were used to identify potential duplicate addresses: 1) address matching based on characteristics of the address derived from MAF data and 2) person matching based on name and date of birth.

The identification of duplicate addresses was carried out in two phases in order to meet the schedule for Accuracy and Coverage Evaluation (A.C.E.). In Phase 1 a provisional list of duplicate addresses was identified. In this phase, address and person matching were carried out independently. Matching addresses were paired. Addresses with one or more exact person matches and similar households were paired. After identifying Kills and addresses given a status of Delete, one address was selected from each remaining pair. The addresses not selected were considered provisional deletions.

No addresses were eliminated from the HCUF as a result of Phase 1 of the duplicate identification process. Phase 2 was implemented after the creation of the HCUF and the results that phase were used to identify which provisional deletions from Phase 1 would be retained on the HCEF. In Phase 2 additional information on address matching and person matching was combined to decide which of the provisional deletions to reinstate. Additional person matching used a modified version of the Census Bureau's probabilistic matching methods. At the completion of Phase 2, a total of 1,392,686 HCUF addresses were identified as duplicate addresses and not retained on the HCEF.

1.2.4 Group Quarters

Most of the information in this section comes from Jonas (2002).

The report covers two aspects of the processing of data for GQ. The first aspect addresses how the population count for GQs was determined and the second addresses the processing of GQ questionnaires with a reported UHE address.

The GQ response data were processed separately from the housing unit data until the final step of the HCUF processing when data from both universes were collected on the same file for the first time.

The Census Bureau relied heavily on the number of GQ questionnaires completed and captured by the DCS2000 to determine the population of each GQ. Individual GQ questionnaires were not tracked during the enumeration processes. Clerical counts of the number of questionnaires at several points of field processing and a count of records by the DCS2000 were recorded. The count of the number of questionnaires was recorded at five points in the post-enumeration processing:

-- By the enumerator immediately following the enumeration of a GQ.
-- By the LCO staff when the questionnaires were received.
-- By the LCO staff when the questionnaires were shipped to the NPC.
-- By the NPC staff when the questionnaires were received.
-- By the DCS2000 during the data capture of questionnaires.

The counts listed above formed the basis for determining the final population count for each GQ. Other processes that contributed data to determining the population count for GQ were:

-- The results of telephone followup interviews with GQ establishments that initially refused to be enumerated. No questionnaires were returned for these GQs. The Census Day population of each GQ was ascertained by the followup interview.
-- Identification of BCF questionnaires with a GQ address.
-- Identification of housing unit questionnaires with a reported UHE address for a GQ.
-- Unduplication of persons at SBE Facilities.
-- Identification of GQ questionnaires with a reported UHE address for a housing unit.

Although residents of all types of GQs were allowed to report UHE (i.e., a Census Day residence other than the GQ at which they were enumerated) only questionnaire data for eligible UHE responses were to be sent to the Non-ID processes. Only persons with eligible UHE responses could be removed from the GQ universe and included in the housing unit universe. Eligibility was determined by the type of GQ from which the questionnaire was received and response to a screening question which identified a person's primary residence.

2. METHODS

The information for this topic report was obtained primarily from the Decennial Management Division (DMD) Operational Assessments and the Census 2000 Evaluation studies that address topics associated with Census 2000 data processing. The major resources for information come from the documents discussed below. For a complete list of resources, see the list of references at the end of the document.

2.1 Decennial Management Division Operational Assessments

The purpose of the operational assessments was to document the successes and lessons learned from the planning and implementation of Census 2000 data processing operations. They provided recommendations for consideration by the research and development, and working groups that will plan similar operations in the future.

The primary source information came from a combination of inputs by all divisions responsible for decennial operations planning and implementation. The recommendations reflect the opinions of the contributors to the assessments and do not necessarily represent the official position of the Census Bureau. The development of the recommendations focused on the individual operation, without an attempt to assess the implications across the entire census process. The DMD operational assessments referenced in this topic report are:

-- Assessment Report for Decennial Response Files DRF1, DRF2, PSA
-- Assessment Report for Non-ID Questionnaire Processing (including BCF/TQA Field Verification)

2.2 Census 2000 Evaluations

Results from the following Census 2000 Evaluation studies contributed to this topic report:

-- Group Quarters Enumeration, E.5
-- Evaluation of Nonresponse Followup - Whole Household Usual Home\Elsewhere Probe, I.2
-- Operational Analysis of the Decennial Response File Linking and Setting of Housing Unit Status and Expected Household Size, L.2
-- Analysis of the Primary Selection Algorithm, L.3.a
-- Resolution of Multiple Census Returns Using a Re-interview, L.3.b
-- Census Unedited File Creation, L.4

2.2.1 Evaluations E.5, L.2, L.3.a, L.4, I.2

These evaluations provide descriptive statistics summarizing the outcome of the Census processes. The results were derived from data files available as products of the Census 2000 processes. These files include:

-- Decennial Response File
-- Hundred Percent Unedited Census File
-- Special Place/Group Quarter Control File
-- Normalized file of matching and geocoding results for enumerator questionnaires from the Non-ID processing

2.2.2 Evaluation L.3.b

The evaluation relies on data from a national sample of addresses affected by the PSA. A post-census interview was conducted at each sample address with someone familiar with the household(s) enumerated on Census questionnaire(s) captured for the address. The residency status of each household member was ascertained through the interview. The data collected were used to judge the accuracy of the decision made by PSA with respect to selection of household members.

3. LIMITATIONS

3.1 General Limitations

The lack of documentation for the Census data processing procedures was an impediment not only for this topic report but for many of the evaluations. The lack of documentation in some cases limited the scope of the evaluations and the details of what we can learn about the processes. It put the evaluations and assessments at risk of conveying false conclusions.

3.2 Evaluation L.2 - Operational Analysis of the Decennial Response File Linking and Setting of Housing Unit Status and Expected Household Size

This evaluation discusses a possible coding error with respect to Simplified Enumerator Questionnaire Interview Summary Item B field on the DRF. The schedule did not allow time to investigate the true extent this error by visually examining the responses on the questionnaire images. The evaluation could only assess the extent of the effect of the error based on related data.

3.3 Evaluation L.4 - Hundred Percent Census Unedited File Creation

An exact tally of the "Kill" addresses was not possible. The DMAF did not provide the information necessary to identify which addresses were identified as a "Kill". As part of a quality assurance on procedures to identify the "Kills", the Decennial Statistical Studies Division (DSSD) implemented software to independently verify the identification. At the time of production the results of the DSSD identification matched exactly the DCSMO production results. The DSSD software was subsequently applied to the current DMAF in order to identify the "Kill" addresses for the evaluation L.4. However, there were approximately 14,000 addresses not identified as a "Kill" that are strongly suspected to be "Kills". This conclusion is based on the data for the results of the field followup operations and the fact that they were not

included the Census. These addresses were assumed to be "Kills" for the purpose of the evaluation.

4. MAJOR FINDINGS

4.1 Decennial Response File Processing - Overall Operational Issues

Much of the information in this section comes from Fowler (2003)

The construction of the Decennial Response File (DRF) is considered a successful element of the Census 2000 data processing operations. The DSCMO successfully collected and processed response data to produce a complete and integrated DRF. The completion of the DRF process was the result of a successful integration of many diverse processing components employed to edit and identify a complete set of response data for each housing unit in the Census 2000.

4.1.1. Requirements Development

Complete documentation of requirements was not developed for the DRF processing or related Quality Assurance (QA) and testing procedures. Requirements were developed for some individual components of the DRF process but not for others.

- Requirements were developed for the process of reformatting data capture input into a standard format (file normalization), the Coverage Edit Followup processing and coding extraction, the linking of continuation forms, the determination of housing status and household size, and the PSA process.

- Requirements were lacking for the other critical steps including interfaces with Non-ID questionnaire processing, and the initial edits of the response data to blank illegal characters, and identifying data defined persons and formatting generational name suffixes.

Late changes to requirements were a challenge for the DSCMO to review and implement. Two examples of changes to requirements that occurred late in the software development cycle were the decision to include Large Households in the Coverage Edit Followup and changes to DRF input data from the Data Capture Audit Resolution processes.

The technical requirements of the questionnaire designs were extensive and the demographic area struggled to finalize questionnaire formats and content within the allotted time frame. A total of 82 different form types were designed for the Census 2000 which resulted in 15 different formats for data capture. The large number of different form types and file formats required greatly increased the schedule and complexity of designing procedures for reformatting the data capture file input into a standard format and for designing the initial edit of the response data.

The lack of timely development of requirements for questionnaire data capture output and the

Non-ID process output hindered the review of documentation and the planning of file testing and QA procedures.

The PSA was the most successfully developed component of the DRF processes. Sufficient staff was dedicated to the development of PSA requirements. The requirements were adequately developed and documented.

4.1.2 Quality Assurance

Appropriate quality assurance safeguards were undertaken and documented in the development of the PSA software development, but were not developed or documented as well for the other components of the DRF processing.

Formal walkthroughs of all PSA specifications were conducted to examine the completeness and identify necessary modifications. The PSA software underwent a very thorough formal testing program. The initial contractor hired to assist in the development of the testing process contributed greatly to the process because he had an abundant knowledge of software testing and a sufficient understanding of the PSA and the DRF process. Midway through the development of software testing the initial contractor was replaced by a contractor who did not have an adequate knowledge of the PSA and the DRF. This became problematic and the remainder of the software testing development was completed by Census Bureau staff. This effort was successful but introduced risks to the success of the testing plan and redirected scarce resources away from other critical operations.

The software testing of other components of the DRF was much more informal. Due to the lack of staff resources and time, there was no formal QA or testing plan for any other component of the DRF. Informal testing was conducted for the processes of linking continuation forms, and determining household status and expected household size. During these steps, problems with related DRF processes were discovered and corrected.

4.1.3 Documentation

Complete documentation of requirements for all DRF components was not accomplished. In some instances it was not evident which staff was primarily responsible for developing requirements. Some of these requirements were not listed on planning schedules and the deliverables were not tracked. The documentation and review of these requirements were not adequate for ensuring their completeness and appropriateness to the task.

4.1.4 Scheduling

The DRF development compensated successfully for late developing data requirements.

Due to scheduling constraints, it was necessary to complete the DRF processing before all inputs were available. More than 207,000 housing unit addresses on DMAF were not included on the DRF. These addresses were added to the DMAF after the DRF1 process merged the data capture file with the DMAF. The housing unit status was subsequently imputed for these addresses in the HCUF process.

4.1.5 Recommendations

These recommendations come from Fowler (2003)

- Design and implement a formal process to develop complete DRF requirements. Planning is needed to ensure sufficient staffing resources are available to adequately plan and develop the requirements.

- Reduce the number of different questionnaire formats to reduce the number of different output formats. This will simplify the data processing.

- Complete the final forms' design much earlier in the planning cycle, desirably prior to the Dress Rehearsal. This will allow sufficient time to develop, document and test specifications for DRF1 processing.

- Improve the planning of software testing and quality assurance (QA) procedures. The testing and QA procedures implemented for the PSA are an example of what we should aim to achieve.

- Reduce the risk associated with placing large responsibilities on a few staff members or one contractor by developing inter-divisional teams. Teams, made up of representatives from various divisions, should develop an early and cooperative partnership. They can ensure that adequate requirements, staffing assessments, and documentation are developed. The structure used for development of the PSA requirements and software is a good example of effective program development.

4.2 Decennial Response File Processing - Phase 2 (DRF2) Processes

4.2.1 Linking Returns

The information in this section comes from Rosenthal (2003)

The DRF included about 1.5 million enumerator continuation returns. All but about 2.3 percent could be linked to a Simplified Enumerator Questionnaire (SEQ). At the completion of the linking process, 33,472 continuation returns remained unlinked to another response record (a

parent record).

Few (126) enumerator continuation forms from List Enumerate (LE) areas were included on the DRF. The DCS 2000 captured many more continuation forms from LE areas but they were not included on the DRF due to an undocumented processing error. The impact of this error on the population coverage is mitigated by fact that the total household size was recorded by enumerators on most SEQs.

The linking process resulted in 1,387,085 DRF returns that were made up of more than one response record.

The linking process looked for SEQs, mail return questionnaires and BCF questionnaires that were used in the place of enumerator continuation forms and attempted to link any of those found to a parent record. Few were found to be used this way.

After the linking process was completed, there were 129,389,529 housing unit returns on the DRF2, excluding returns that were ineligible to receive a status. Returns that were ineligible for status assignment include 197,091 blank returns and 696,691 replaced SEQ returns. Replaced SEQ returns are returns replaced by another SEQ as a result of a field operations quality assurance check of enumerators' work. Table 1 below shows the number of eligible returns after the linking process was completed. For the results shown in this table, addresses are grouped by the number of response records that contributed to a return.

Table 1. Number of Response Records Comprising a Return

Response Records Per Return	Number of Returns	Percent
1 *(No continuation forms)*	128,002,444	98.9
2 *(1+ continuation forms)*	1,347,977	1.04
3 or more *(2+ continuation forms)*	39,108	0.03
Total Returns	129,389,529	100.00

Source: Rosenthal (2003) Table 4

4.2.2 Assigning Housing Unit Status and Household Size

Table 2 below shows the housing unit statuses assigned to the 129,389,529 returns on DRF.

Table 2. Housing Unit Status by Type of Return

Housing Unit Status	Type of Return Number (Column Percent)			
	Total	**Self Response**	**Enumerator Response**	**GQ/BCF**
Occupied	112,150,512 (86.7)	81,080,662 (100.0)	30,465,137 (64.7)	604,713 (50.0)
Vacant	14,141,843 (10.9)	18,504 (0.0)	14,123,339 (30.0)	0 (0.0)
Delete	1,778,824 (1.4)	0 (0.0)	1,778,824 (3.8)	0 (0.0)
Occupied/Unresolved Household Size	934,849 (0.7)	0 (0.0)	329,895 (0.7)	604,954 (50.0)
Occupancy Status Unknown	329,804 (0.3)	538 (0.0)	329,266 (0.7)	0 (0.0)
Status Undetermined	53,697 (0.0)	0 (0.0)	53,697 (0.0)	0 (0.0)
Total	129,389,529	81,099,704	47,080,158	1,209,667

Source: Rosenthal (2003)

- About 62.7 percent of the returns were self response returns which includes all mail back questionnaire returns, internet responses and responses received through the TQA operation.

- About 36.4 percent of the returns were SEQ returns.

- The DRF2 also included 604,954 returns for individuals enumerated at a GQ claiming a usual home elsewhere in the housing unit universe (0.5 percent of the DRF returns), and 604,713 BCF returns (0.5 percent of the DRF returns). The GQ returns were all assigned a status of Occupied/Unresolved Household Size and the BCF returns were all assigned a status of Occupied.

4.2.2.1 Unresolved Housing Unit Status Among Enumerator Response Returns

The key data items from enumerator returns used to assign the housing unit status are the following:

-- The number of persons enumerated on the questionnaire
-- The value of the respondent reported household size
-- The response to the SEQ Interview Summary Item A (Status on April 1, 2000)
-- The response to the SEQ Interview Summary Item B (POP on April 1, 2000)
-- The response to the SEQ Interview Summary Item C (Type of vacant unit)

Occupied/Unresolved Household Size

On most of the 329,895 enumerator returns assigned the Occupied/Unresolved Household Size status, the enumerator clearly indicated that the unit was occupied but that the household size could not be determined. Almost 72 percent of these returns had an Interview Summary Item A value of 'Occupied' and an Interview Summary Item B value of 'POP Unknown'.

Responses on nearly all of the remaining returns clearly indicated that the housing unit was occupied but the return lacked sufficient data on the size of the household. Almost all (98 percent) of these returns had an Interview Summary Item A value of 'Occupied' and an Interview Summary Item B value of 1 to 97, but there were no persons enumerated on the questionnaire and the respondent reported household size was 0 or missing. The responses to Interview Summary Item B were numeric values captured by an optical character recognition methodology. This method can introduce error in the response captured, there were no data that confirmed the household size captured for the Interview Summary Item B.

Occupancy Status Unknown -

The responses to as many as 258,963 (78.5 percent) of the returns given a housing status of Occupancy Status Unknown may have been incorrectly coded in the DRF2 during a data editing process. Responses of '0' to Interviewer Summary Item B were incorrectly recoded as 'missing' by the edits of the DRF2 data. This error caused many returns to be erroneously given a status of Occupancy Status Unknown instead of a status of Vacant.

It is impossible to know for sure how many of the 258,963 were affected by the coding error. However, there is convincing evidence that nearly all of them were affected by the coding error. More than 94 percent of these returns had a response to the Interview Summary Item C which allowed the enumerator to report the type of vacancy of the address. Interview Summary Item C was only filled for vacant units. This suggests that enumerators took care to fill Interview Summary Item B with '0' as well as filling Interview Summary Item C for almost all of these cases.

The DRF2 returns potentially affected by the coding error were a large portion of the addresses placed in the Occupancy Imputation category during the HCUF processing and given an imputed housing unit status. They accounted for about 74 percent of the 195,245 housing units on the HCUF placed in the Occupancy Imputation category. As such, the coding error contributed to an undercount of vacant housing units and an over count of occupied housing units.

Status Undetermined

Almost 91 percent of the 53,697 returns given a housing status of Status Undetermined had no persons enumerated on the questionnaire and an Interviewer Summary Item A value of 'Delete', but they had a conflicting value in the Interviewer Summary Item B. The Interview Summary Item B response for these cases was 1 to 97, or Pop Unknown.

4.2.2.2 Resolution of Housing Unit Status by NRFU vs. CIFU

About 2.3 percent of the enumerator returns received from the CIFU were assigned one of the three unresolved housing unit statuses discussed above, while only 1.34 percent of the NRFU enumerator returns had one of these statuses.

The rate that enumerator returns were assigned an Occupied/Unresolved Household Size status was twice as large for CIFU compare to NRFU, 1.3 percent vs. 0.6 percent, respectively.

There were more than 4.8 million addresses visited in both the NRFU and CIFU operations. However, data on only 4,233 of these addresses visited in both operations were so insufficient that they resulted in the assignment of an unresolved housing unit status. This shows that a census enumerator completed an enumeration at all but small number of housing units included in the NRFU.

4.2.2.3 Proxy Responses for Enumerator Returns

The respondents for about 17.4 percent of the occupied enumerator returns were proxy respondents (i.e., the respondent did not belong to the household enumerated on the return). This rate does not include cases for which the type of respondent is unknown. Vacant and Delete returns are, by default, all said to have a proxy respondent because there is no household respondent in these cases.

The returns given an Occupied/Unresolved Household Size housing unit status had a much higher rate of proxy respondents (30.8 percent) as would be expected.

The returns in the two other unresolved housing unit status categories (Occupancy Status Unknown, Status Undetermined) had a very high rate of proxy respondents. About 76.5 percent of the respondents for these returns were proxy respondents. This high rate is misleading because more than 67 percent of these returns may have been erroneously given an Occupancy Status Unknown instead of Vacant, as noted in an earlier discussion. Respondents for vacant

housing units are expected to be proxy respondents, thus the high rate of proxy respondents for these cases.

4.2.2.4 Setting of Expected Household Size

There was a high level of consistency on each return among the key data items used to assign the expected household size. Among both mail and enumerator returns there was agreement between keys items for over 93 percent of the returns. On mail returns the key items were the respondent provided household size and the number of person enumerated on the return. The key items on enumerator returns were the Interviewer Summary Item B (POP on April 1, 2000) and the number of persons enumerated on the return.

4.2.3 Primary Selection Algorithm

Most of these results come from Baumgardner (2002).

The PSA was applied to 127,610,705 eligible returns at 118,360,443 census addresses.

- About 7.6 percent of the addresses on the DRF2 had two or more returns with 7.4 percent of all addresses having just two returns.

- There were another 158,530 addresses that had only returns not eligible for the PSA. The returns ineligible for the PSA included blank returns, those assigned the status of Delete and enumerator returns that are unusable because they were replaced by another SEQ as a result of a quality assurance check of enumerators' work.

Formation and Selection PSA Households at Addresses with Two or More Returns

A total of 11,426,952 PSA households were formed at 8,960,245 addresses with two or more eligible returns.

Only one PSA Household was formed at 6,564,116 (73.3 percent) of these addresses.

- About 40.4 percent of these PSA Households were vacant housing units.

Two or more PSA Households were formed at 2,396,129 addresses.

- These addresses make up just 2.0 percent of all DRF2 addresses.

- Three or more PSA Households were formed at just 46,141 addresses.

- A total of 1,235,327 addresses (51.6 percent) had one vacant PSA Household and one or more non-vacant PSA Household. The PSA selected the vacant PSA household

over the non-vacant household(s) at only 194,596 of these addresses.

In order to identify the Primary PSA Household, the PSA applied, sequentially, an ordered list of seven criteria. Two criteria caused a return with a resolved housing unit status to be selected over returns with an unresolved housing unit status (POP Count Status criterion) and selected the return that was the highest in a hierarchy of questionnaire form types (Return Type criterion), respectively.

- The Return Type criterion was the criterion that selected the Primary Household about 74 percent of the time.
- The POP Count Status criterion was the selection criterion about 16 percent of the time.
- No other criterion triggered the selection of more than 3.2 percent of the Primary Households.

Formation of Primary Selection Algorithm Households at Addresses with Two Returns

About 97.3 percent of the addresses with two or more returns had just two eligible returns. Among these 8,716,359 addresses, the two returns were combined to form one PSA household 74 percent of the time.

Table 3 shows the results of forming PSA households at those addresses with just two eligible returns.

Table 3. Formation of PSA Households at Addresses with Two Returns

Outcome of PSA Household Formation	Number	Percent of Total
One PSA Household Formed	6,463756	74.1
Both returns are Vacant	2,634,322	30.2
The Basic Return contains all of the persons on the other return	3,469,789	38.8
The returns have person(s) in common but the Basic Return does not contain all of the persons on the other return	359,645	4.1
Two PSA Households Formed	2,252,603	25.8
One Non-Vacant[4] and One Vacant	1,162,675	13.3
Both Non-Vacant	1,089,928	12.5
TOTAL	8,716,359	

Source: Baumgardner (2003) Table 14

There were 3,829,434 addresses at which only one occupied PSA household was formed.

- Only about 9 percent of these are cases where the Basic Return did not contain all of the persons enumerated on the other return.
  - An estimated 82 percent of these PSA Households were correctly formed, i.e., each of the returns that made up the household had at least one census resident.

[4] The term Non-Vacant includes Occupied returns and returns with an unresolved housing unit status.

There are 1,089,928 addresses where two returns for an occupied household could not be combined into one PSA household.

- At an estimated 38 percent of the addresses with two occupied returns and two PSA Households, both returns represented the same household. That is, there were residents in both returns.
  - The PSA had no chance of combining an estimated 75.1 percent of these by matching persons because there were no duplicate persons between the two forms.
  - The PSA performed matching and failed to identify duplicate persons at an estimated 16 percent of these addresses.
- At an estimated 58 percent of the addresses with two occupied returns and two PSA Households there were census residents on only one return.
  - PSA chose the correct PSA Household in about 65 percent of these cases.

Primary Selection Algorithm Household Formed from Two Returns

A total of 6,561,984 PSA Households (57.4 percent of all PSA Households) were comprised of two returns. Table 4. below shows the combination of return types for these PSA households.

Table 4. Combinations of Return Types for PSA Households with Two Returns

Combination of Returns	Number of Addresses	Percent of Addresses
Mail/Mail	196,751	3.0
Mail/Enumerator	2,732,392	41.6
Enumerator/Enumerator	2,845,843	43.4
Mail/RPA[5] & Enumerator /RPA	782,906	11.9
RPA/RPA	4,092	0.1
Total	6,561,984	

Source: Baumgardner (2003) Table 9

[5]RPA (Respondent Provided Address) returns -These include GQ returns with a Usual Home Elsewhere housing unit address, BCF returns, NRFU returns for a Whole Household Usual Elsewhere and In-Mover address.

The large proportion of PSA households that contain two enumerator returns is primarily the result of the CIFU operation design. Most addresses that were determined to be vacant by the NRFU were included in the CIFU. Whenever CIFU determined that one of these as occupied or vacant at least two enumerator returns were captured for the census address.

The large proportion of PSA households made up of a mail and an enumerator return is primarily due to mail returns that were received after the identification of the NRFU universe.

Only a small proportion (14 percent) of the PSA households with two mail returns is the result of a request for a foreign language questionnaire. Mail returns include paper mail back returns, internet responses and TQA reverse-Computer Assisted Telephone Interview (CATI) responses. It appears that the remainder of these PSA households represent duplicate attempts by respondents to report their households using two of these three methods.

The PSA automatically individuals from some returns for an RPA address to the Primary PSA household. Evaluation L.3.b. estimated that at least 60 percent of the individuals added from RPAs in this fashion were correctly included in the Census.

4.2.4 Recommendations

These recommendations come from Rosenthal (2003) and Baumgardner (2002)

4.2.4.1 Linking Returns

Attempt to link only enumerator questionnaires and enumerator continuation forms, if these forms are used in the future. Doing so would greatly simplify the requirements of the linking process while causing negligible loss of data and possibly no effect on population count.

Ensure that all continuation forms are included on the DRF.

4.2.4.2 Determining Housing Unit Status and Household Size

Construct more comprehensives instructions for enumerators to assist them in completing questionnaires for complicated and unusual interviews.

Pursue the use of computer assisted personal interviews through the use of hand held computer devices.

4.2.4.3 Primary Selection Algorithm

Define a simpler PSA process that relies more on type of return, source of return, and status of return and less on person matching. Person matching added much complexity to the process but affected the selection of a Primary PSA household at only a very small

number of addresses.

Plan for the PSA to address only those combinations of returns that can be predicted by the design of the census operations. The PSA was robust and was designed to handle a large variety of cases including combinations of returns not anticipated by the design of census operations. Few combinations of returns occurred that were not anticipated by the census design.

Conduct research on the feasibility of integrating the PSA with processes to identify duplicate addresses and persons duplicated at more than one address.

Pursue methods to reduce the inclusion of mail response households in the NRFU. Mail returns were received for more than 3.4 million addresses after the identification of the NRFU universe. As a result, the DRF had both a mail return and an enumerator questionnaire for these addresses.

4.3 Hundred Percent Census Unedited File for Housing Units

Most of these results come from Jonas (2003).

4.3.1 The Hundred Percent Census Unedited File Processing

The first step in the Hundred Percent Census Unedited File (HCUF) processing for housing unit data was to identify addresses on the DMAF that did not represent a census housing unit. The addresses eliminated from the housing unit universe at this stage of processing were referred to as 'Kills'. 'Kills' were almost entirely address location confirmed not to be housing units by the many address develop operations in the census.

- There are 127,828,778 addresses on the DMAF of which 9,057,195 (7.1 percent) were identified as a "Kill".

The DMAF addresses that remained in-scope after the 'Kills' were identified were merged with the DRF2. At this stage of processing one of the following housing unit statuses was assigned to each address on the DMAF: Occupied, Vacant, Delete, Occupied with Unresolved Household Size, Occupancy Status Unknown, and Status Unknown. These statuses correspond to those assigned to individual census returns during the DRF2 processing.

Table 5 shows the final status assigned to each addresses that was in-scope at this stage.

Table 5. Source of Housing Unit Status for DMAF Addresses

	Source of Status Data					
Housing Unit Status	Self Response	Enumerator Response	Respondent Provided Response	No Data	Total	%
Resolved Occupancy Status:						
Occupied	80,781,126	26,636,881	197,778	0	107,615,785	90.6
Vacant	16,277	10,438,871	0	0	10,455,148	8.8
Delete	0	8,653	0	1	8,654	0.0
Occupied/Unresolved Household Size (Household Size Imputation)	0	169,902	30,232	0	200,134	0.2
Unresolved Occupancy Status:						
Occupancy Status Unknown (Occupancy Imputation)	506	194,739	0	0	195,245	0.2
Status Undetermined (Status Imputation)	0	45,113	27	251,477	296,617	0.2
Total	80,797,909	37,494,159	228,037	251,478	118,771,583	100.0
%	68.0	31.6	0.2	0.2	100.0	

Source: Jonas (2003) Table 2

The source of the status for each address could be either the status assigned to the DRF return selected by the PSA or the DMAF data on the outcome of the NRFU, CIFU or Field Verification (FV) operations. The categories for the source of the data on which the housing unit status was assigned are defined below:

-- Self Response - the source was a DRF2 return for a paper mail return questionnaire, internet response or a reverse CATI response.

-- Enumerator Response - the source was a DRF2 return for a Simple Enumerator Questionnaire return, an enumerator continuation form, or the DMAF data on the outcome of the NRFU, CIFU or FV.

-- Respondent Provided Address - the source was a DRF2 return for a paper BCF return or a GQ return.

-- No Data - This source indicates that there was no return on the DRF2 for the address, and that there were no data on the DMAF from the NRFU, CIFU or FV operations.

During the imputation of housing unit status to addresses in the Status Imputation category, a total of 47,126 addresses were given a housing unit status of Delete and removed from the HCUF.

No Data for a Housing Unit - There were no data for almost 85 percent of the addresses assigned a housing unit status of Status Unknown.

- Over 82 percent of these are addresses added to the DMAF from updates that occurred in August 2000 or later.

  Those addresses added late in the processing schedule were added to the DMAF after the merge of the DMAF and DRF2 data. Thus, the data capture responses for these addresses were not included on the DRF2. None of these were addresses pre-assigned to the NRFU, CIFU or FV operations although a large proportion of these were added to the Census during these operations. Since the results of these followup operations were only recorded on the DMAF for addresses pre-assigned to the operation, no data on housing unit status are available on the DMAF for these addresses.

4.3.2 Duplicate Delete Processing

The Duplicate Delete process identified 1,392,686 duplicate housing units that were deleted at the time the HCEF creation.

- About 2.9 percent of the deleted housing units had a Vacant housing unit status.
- About 0.6 percent of the deleted housing units had a pre-imputation housing unit status of Occupancy Status Unknown or Status Unknown.

4.3.3 Recommendation

- Reexamine the timing and coordination of data processing operations in order to ensure that the responses captured for all addresses can be included in the final census files. (Jonas, 2003)

4.4 Non-ID Addresses Processing

Most of these results come from Medina (2001).

4.4.1 Non-ID Processing Results

The geocoding and matching operations were effective and efficient processes.

- The clerical geocoding operation utilizing the Interactive Mapping and Geocoding System, an interactive clerical matching and geocoding operation which involves calling respondents while allowing clerical staff to simultaneously see both the MAF

and TIGER databases, is a viable and effective operation.

- The clerical staff was able to match and geocode addresses at a much faster rate than estimated. These faster production rates significantly contributed to lower staffing levels than planned for this clerical operation.

- The automated non-city-style matching was an effective tool for reducing the workload for clerical matching and geocoding. Based on the clerical geocoding rate, the matching saved approximately 300 person days clerical staff time.

The workload of Type A and Type B records for the Non-ID process was more than two times as large as it should have been. Almost 2.3 million of the 4.2 million Type A and Type B Non-ID cases were included in error. The DCSMO did not apply the filter to exclude ineligible GQ UHE returns from the Non-ID process prior to identifying returns that required the assignment of a MAFID through the Non-ID process. As a result, 2,281,712 GQ returns were erroneously included in the Non-ID process while only 659,566 GQ returns were legitimately included.

The GEO received more than one million records too late to be appropriately processed in subsequent Census 2000 operations. Although we do not have direct measurements of the errors caused by the tardy transfer of records, an examination of how these records were treated in Census processes illustrates the potential for serious coverage errors and loss of data.

- The DCSMO delivered over 830,000 Type A and Type B records to the GEO after the June 14, 2000 processing cut-off date for identifying Type A addresses that should be included in the FV operation. The records were received too late for the Non-ID process to complete them prior to the deadline for identifying the FV universe.

  Any of these records for eligible GQ UHE addresses that could be geocoded but not matched to the MAF were eligible for the FV operation. Since they were processed too late to be included in the FV, they were added to the DMAF without verification.

- More than 78,000 such addresses were added to the Census without having been included in the FV operations. Most of these addresses were obtained during the NRFU interview through the UHE and In-Mover probes.

  - Most of these address obtained through the NRFU UHE probe. A total of approximately 55,000 addresses obtain through this probe should have been sent to FV. Only one percent of these were sent to the FV. The remainder were added to the Census without verification. (Viator, 2003)

    Most of the added addresses obtained through the NRFU UHE probe are suspicious additions to the Census 2000. More than 70 percent of the approximately 54,000 addresses obtained through this UHE probe were reported to be vacant on April 1, 2000. A report of a vacant housing unit is

contrary to the concept of the usual home elsewhere.

Furthermore, the FV operation found that about half of the UHE addresses that were included in the field operation were not housing units. This rate is consistent with the overall FV operation which found about only one-half of the cases processed were census housing units.

- The GEO received over 6,800 Type A and Type B records from the DSCMO after the August 4, 2000 processing cut-off date for the August 15, 2000 delivery of MAF updates to the DMAF. Response for addresses added to the MAF after this update were not included on the DRF.

- The GEO received more than 124,000 UHE addresses for Individual Census Questionnaires (ICQs) and Shipboard Census Reports (SCRs) after the July 23, 2000 suspension of the clerical geocoding processing. As a result, no attempt was made to clerically geocode any of these addresses that failed the automated matching and geocoding process. Thus, new addresses could not be identified and included in the Census.

- More than 207,000 Type C records were processed too late for their response data to be included on the DRF or included in the FV process. All of the Type C addresses identified as having been processed too late were addresses that were new to the MAF. The DMAF was updated with these new addresses after the DRF2 was created.

4.4.2 Recommendation

- Continue to refine and test the Interactive Matching and Geocoding System (IMAGS) software as a product to clerically match and geocode addresses. (Medina, 2001)

4.5 Group Quarters Processing

Most of these results come from Jonas (2002)

4.5.1 Resolution of Missing Data

The GQ processing successfully dealt with difficulties surrounding a potentially large amount of missing data. In May 2000, the NPC reported that a large number of GQ questionnaires did not have GQ Identification (ID) numbers on them and/or had no associated control sheet. Procedures were quickly designed and implemented by Census Bureau Headquarters staff to clerically review these questionnaires and, if possible, identify them with the correct GQ. An estimated 700,000 questionnaires were reviewed during this operation. This operation appears to have been highly successful although no official accounting was made of the outcome of this review.

A unique barcode and number were printed on each GQ questionnaire in the Census 2000. However, the barcode was not used to track GQ questionnaires from enumeration through data capture. Census enumerators were required to transcribe the 14 digit GQ identification number to each GQ questionnaire. When this was not done or was done incorrectly it was difficult and sometimes impossible to identify the GQ at which the respondent was enumerated.

After GQ questionnaires were captured by the DCS2000, the counts of captured questionnaires were examined by an inter-divisional team of staff knowledgeable of the GQ enumeration and processing operations. This review was not originally part of the design for GQ processing. The team found that the data capture was incomplete in several ways:

-- No questionnaires were received for a number of GQs which were believed to have refused our attempts to enumerate them.

-- The count of questionnaires for a number of GQs was far less than projected by pre-enumeration operations.

-- A number of GQs had a higher count of questionnaires sent to NPC by the LCOs than were captured by the DCS2000.

(A portion of the missing questionnaires can be attributed to the missing GQ ID numbers on some forms and our inability to identify them with the appropriate GQ.)

A previously unplanned telephone followup operation was implemented to address the first two of the count deficiencies described above. This followup ascertained the Census Day population count for GQs but did not collect the demographic data of residents.

- A total count of 101,598 persons was added to the GQ population as result of this followup. This was 1.3 percent of the total Census 2000 GQ population.

- About 4.4 percent of GQ residents at hospitals were enumerated by this followup.

The DSSD designed a procedure to derive a count of the expected number of persons enumerated at GQs to mitigate the problems posed by the last of the three count discrepancies. When the aggregate count of forms shipped to the NPC for a Special Place was higher than the aggregate count of forms captured, the difference in these two counts was allocated to the GQs within the Special Place proportional to the differences in the two counts for each GQ.

Collectively, all these operations added about 200,000 persons to the Census 2000 GQ population of 7,825,407. As a result, it was necessary to impute demographic data for 2.6 percent of the Census 2000 GQ population.

4.5.2 Processing of Responses with a Usual Home Elsewhere Address

The GQ processing successfully recovered from the erroneous routing of returns for GQ residents reporting UHE addresses to the Non-ID process. The GQ responses sent to the Non-ID process could be removed from the GQ universe and placed in the housing unit universe if the UHE address was confirmed to be a housing unit so it was important that we successfully identified ineligible GQ UHE responses thereby preventing them from being erroneously removed from the GQ universe.

- A total of 659,566 responses with a UHE housing unit address were correctly removed from the GQ universe.

- A total of 150,315 responses were incorrectly removed from the GQ universe because they were incorrectly identified as having a UHE address.

- GQ processing erroneously sent nearly 2.3 million GQ responses to the Non-ID process.

    -- There were 1,892,742 responses with a UHE address collected from those types of GQs that made them ineligible to be sent to the Non-ID process.

    -- There were 388,970 responses that were incorrectly identified as having a UHE address.

4.5.3 Recommendations

These recommendations come from Jonas (2002).

- Track GQ questionnaires throughout the operation, from enumeration through data capture. A unique barcode and number printed on each GQ questionnaire can be used for this purpose using existing products to record and manage a database of these identification numbers.

- Institute more effective software quality assurance programs involving representatives from more than one division.

5. CONCLUSIONS

5.1 Processing Systems Design Architecture

- No critical failures of the processing system design were reported during the implementation of the census. The design of the census processing systems for housing unit responses was adequate for the required tasks. The processing systems handled millions of census responses from a large variety of data collection methods and data

collection systems. Enumeration results, demographic data, housing unit data, and geographic information were successfully integrated, on time, into the critical Census databases such as the DRF, the HCUF and the DMAF.

5.2 Development and Documentation of Requirements

- A common thread in all of the studies was the need for well documented processing requirements. It has been suggested that more time be allocated to the development and design of requirements.

  According to Fowler (2003) the requirements development for the DRF1 and DRF2 were lacking. Complete integrated requirements development was not available to address all components as necessary to produce adequate DRF1 and DRF2 documentation, and design quality assurance processes and software testing. Instead, requirement documents were produced piecemeal, which resulted in processing complications.

  The DRF1 requirements documentation existed for some components of the file creation process, such as creation of the normalized data response files, the interface with the Coverage Edit Followup (CEFU) operation and coding extraction, but did not exist for other critical steps such as interfaces with the Non-ID process, and the editing and coding of the response data. Late changes to requirements to address the inclusion of large households in the CEFU and changes to the Data Capture Audit resolution process were challenging for DCSMO to implement. There are no documented software testing or QA processes for these operations.

  The steps undertaken to develop and implement requirements for the Primary Selection Algorithm (PSA) was an exemplary and successful process. Considerable staff resources were devoted to the development of requirements and software for the PSA which was applied to less than ten percent of the Census 2000 housing unit addresses. As stated by Fowler (2003), the development of requirements for the PSA was adequate to the task. The dedication of sufficient staff to this task contributed to the successful development of requirements. The requirements for the PSA software consisted of a very complex set of criteria and person matching. The timely development and documentation of these requirements allowed for the design and implementation of more complete and effective software testing and QA procedures.

  Processing requirements were also fully developed for the processes of linking questionnaire and continuation form response data, and assigning housing unit status and household size. However, the software testing for these processes was much more informal than for the PSA. Insufficient resources and the lateness of requirements development did not allow adequate time to develop formal software testing and QA procedures.

- A complete identification of the requirements needed was not accomplished prior to the implementation of census operations. Some requirements documents were not listed in

the Master Activity Schedule. The responsibility for these requirement specifications was not assigned with sufficient time to complete all steps of a full process development and the deliverables could not be tracked.

There are several examples of processes for which the need for requirements and for which the assignment of the responsibility for specifying requirements was not identified in a timely manner. These include the DRF2 processing (requirements for linking response records and requirements for assigning housing unit status), and the CUF processing (requirements for identifying 'Kills', integrating of the DMAF and DRF2).

5.3 Quality Assurance Processes

- The lack of quality process control and quality assurance software testing put many of the processing steps at risk. The overall success of the Census 2000 processing is apparent from the studies that contributed to this report. However, some missteps did occur that had important effects on Census data. These may have been preventable through more formal and thorough quality assurance and control procedures.

- A well designed and formal QA program was carried out for the PSA. The PSA used inter-divisional teams to develop requirements and software testing procedures. Formal walkthrough and testing were conducted for all software components.

  Primary responsibility for the design of testing procedures for the PSA software was initially given to contractors. This worked well in the beginning because the contractor had expertise in software testing and had gained sufficient knowledge of the PSA process through the Census 2000 Dress Rehearsal experience. Midway through the development process, this contractor left and was replaced by others who did not have an adequate understanding of the PSA software requirements. From that point on, it became necessary for Census Bureau staff from the DSSD to assume primary responsibility for the design of the software testing. By this stage of development the Census Bureau staff had gained sufficient knowledge of the principles of software testing. The software testing was successful but the transition of responsibilities put the testing process at great risk.

5.4 Non-ID Process

- The geocoding and matching operations of the Non-ID process were effective and efficient processes, yet Non-ID processing was not completed in time so that the data form all cases could be integrated into the DRF2 and the HCUF.

  It appears that the Non-ID process was overwhelmed with the enormous number of GQ UHE addresses erroneously included the process by the DCSMO. Jonas (2002) reported that the GQ processing operation erroneously sent nearly 2.3 million UHE addresses to the Non-ID process. This was over half of the Non-ID workload. It appears that due to this large workload, many Non-ID records were delivered to the GEO after important

processing cutoff dates (Medina, 2001). Medina (2001) points out that the GEO received more than 830,000 Type A and B addresses from the DCSMO later than the June 14, 2000 processing cutoff date for the identification of the FV workload. Many of these were delivered to the GEO, after the a cutoff for inclusion of the response data into the DRF2. Although Medina (2001) does not discuss Type C addresses in the Non-ID processes, the late delivery of such a large number of Type A and B records almost certainly delayed the processing of Type C records.

The need to adhere to a tight processing schedule meant that other processing took priority over the processing of Non-ID records by the DSCMO. All Non-ID records were ultimately delivered by the DCSMO to the GEO, processed by the GEO and returned to the DSCMO processing queue. All Non-ID addresses geocoded by the GEO were included in the DMAF.

The completion of Non-ID processing late in the Census schedule had two important impacts on the Census enumerations. The first is that the response data from more than 207,000 housing units were not included in the Census. Some of these housing units were deleted from the Census by the count imputation process. The second is that it may have added many nonexistent housing units to the Census. The late timing of Non-ID processing meant that more than 78,000 housing units were added to the census without having been verified by the FV operation. An evaluation of the FV found that only about half of the addresses processed by the FV were verified as valid housing units.

5.5 Count Imputation of Housing Unit Status

- It was noted by Fay (2001) that the Census 2000 experienced a higher rate of whole person imputations than the 1990 Census. The count imputation process accounted for 0.42 percent of the total Census 2000 population, a rate several times higher than experienced in the 1990 Census.

Count imputation included the three categories of whole household imputation in which a housing unit status and/or household size were imputed to census addresses. These three categories include a total of 691,996 addresses: 1) 200,134 for Household Size Imputation, 2) 195,245 for Occupancy Imputation and 3) 296,617 for Status Imputation.

It appears that processing errors caused the missing data for a significant portion of the addresses in the latter two imputation categories. These errors may have tripled the number of Census housing units for which the occupancy status was imputed. How the number of addresses in each of these two categories was affected by deviations from specifications is discussed below:

Status Imputation - The Status Imputation category includes a total of 251,477 addresses for which there was no return on the DRF. More than 80 percent (207,283) of these addresses have been identified as Type C addresses processed very late by the Non-ID process and added to the DMAF in August 2000 or later. These were part of the updates

to the DMAF that occurred as late as November 2000. These updates to the DMAF occurred so late that the response data for these added addresses could not be included on the DRF2. As a result it was necessary to impute the housing unit status and household size for these addresses.

Occupancy Imputation - The Occupancy Imputation category includes 145,367 Census addresses which are suspected to be vacant housing units but were included in this imputation category. Rosenthal (2003) states that responses of '0' to the Interviewer Summary Item B (POP on April 1, 2000) were mistakenly coded as a blank entry. This processing error occurred on as many as 258,963 returns assigned the housing status of Occupancy Status Unknown. Almost 95 percent of these returns had the Interview Summary Item C (Type of Vacancy) filled. Enumerators were instructed to fill Interview Summary Item C only for vacant housing units. This suggests that enumerators took care to enter '0' in Interview Summary Item B as well as filling Interview Summary Item C for almost all of the 258,963 responses with a suspected error. The PSA chose these returns as the Primary PSA Household at 145,367 addresses included on the HCUF before the imputation of housing unit status.

5.6 Primary Selection Algorithm

- For addresses with two returns, the outcome of the PSA could have differed for fewer than an estimated 500,000 addresses had it not included a within address person matching function. This number includes an estimated 104,000 addresses with two returns for the same household that the PSA failed to match. This does not imply that the results would have necessarily been different or that they would have been less correct.

  More than ninety-seven of all addresses with multiple returns had just two returns. One of the following situations exists for all of the addresses with two returns: 1) all Vacant returns, 2) one Vacant return and one Occupied return, 3) a Basic Return that included all of the persons on the other return, 4) two returns for two different households, or 5) two returns with matching persons and each having unique persons not found on the other return. Only for the last scenario could the person matching have affected the accuracy of enumerations resulting from the PSA outcome.

6. RECOMMENDATIONS

- Ensure timely and complete documentation of processing requirements and design. Timely processing requirements will reduce the time required for software development, allow for adequate software testing, and allow for the design and implementation of quality assurance processes. The timely development of documentation will allow for complete and accurate assessments of the interdependencies between procedures. Preferably, complete documentation would be achieved for a Census dress rehearsal. In most cases the final Census documentation would evolve directly from the dress rehearsal documentation reflecting small changes based on lessons learned in a dress rehearsal. Having full documentation completed for the dress rehearsal would ensure that the impact of modifications to one procedure on related activities could be adequately evaluated.

- Identify the staffs with the critical skills and knowledge needed to develop all requirements early in the development process.

- Ensure that effective QA procedures are in place for all census data processing operations. Allow flexibility in the standards on which the QA procedures are based. The scope of QA procedures and resources required to implement them should be commensurate with the level of risks associated with processing errors.

- Incorporate the use of interactive geocoding software such as the Interactive Matching and Geocoding Systems (IMAGS) software again for the 2010 Census. Continue to develop and test software such as this as a product to clerically match and geocode addresses. The use of this software resulted in an efficient clerical geocoding operation with respect to timing and outcome. The process took much less time than expected.

  The addresses geocoded using this software were addresses that could not be geocoded by an automated process. Yet, the proportion of clerically geocoded addresses found by FV in the block to which they were coded was similar to the proportion for addresses that could be geocoded by the completely automated process. This indicates that an adequate level of accuracy was achieved for the clerical geocoding process using the interactive geocoding software.

- Eliminate the within address person matching function from the PSA. Define a simpler within address return selection process that relies more on type of return and status of return and less on person matching. In the Census 2000, the elimination of within address person matching would not have significantly degraded coverage or the quality of Census data. Extensive resources in the Census 2000 were devoted to designing and implementing a PSA that relied on within address person matching. In the 2010 Census, these resources may be better directed toward the unduplication of persons across addresses.

REFERENCES

Baumgardner, Stephanie (2002), *Analysis of the Primary Selection Algorithm*, Census 2000 Evaluation L.3.a., November 2002.

Baumgardner, Stephanie (2003), *Resolution of Multiple Census Returns Using a Re-interview*, Census 2000 Evaluation L.3.b., September 2003.

Carter, Nathan (2002), Be Counted Campaign for Census 2000, Census 2000 Evaluation A.3, September 2002.

Coan, Edmund (2000), *Program Master Plan: Census 2000 Decennial Response File Program*, Census 2000 Informational Memorandum No. 85, December 2000.

Fay, Robert (2001), *The 2000 Housing Unit Duplication Operation and Their Effect on the Accuracy of the Population Count*, Joint Statistical Meetings Proceedings 2001, August 2001.

Fowler, Charles (2003), *Assessment Report for Decennial Response Files DRF1, DRF2, PSA*, Census 2000 Informational Memorandum No. 137, May 2003.

Griffin, Richard (2001), Census 2000 - *Missing Housing Unit Status and Population Data*, DSSD Census 2000 Procedures and Operations Memorandum Series B-17, February 2001.

Jonas, Kim (2002), *Group Quarters Enumeration*, Census 2000 Evaluation E.5, September 2002.

Jonas, Kim (2003), *Census Unedited File Creation*, Census 2000 Evaluation L.4, August, 2003.

Medina, Karen (2001), Assessment Report for Non-ID Questionnaire Processing (including BCF/TQA Field Verification), September 2003.

Nash, Fay (2001), *Analysis of Census Imputations, Executive Steering Committee For A.C.E. Policy II Report No. 21*, September 2001.

Pike, A. Edward (2001), *Program Master Plan: Census 2000 Matching/Geocoding Non-Master Address File (MAF) Identification (ID) Questionnaires*, Census 2000 Informational Memorandum No. 98, April 2001.

Rosenthal, Miriam (2003), *Operational Analysis of the Decennial Response File Linking and Setting of Housing Unit Status and Expected Household Size*, Census 2000 Evaluation L.2, June, 2003.

Tenebaum, Michael (2001), *Assessment of Field Verification*, Census 2000 Evaluation H.2, July 2001.

Viator, Mark; Alberti, Nicholas (2003), *Evaluation of Nonresponse Followup - Whole Household Usual Home Elsewhere Probe*, Census 2000 Evaluation I.2, February 2003.